Prior Interaction, Identity, and Cooperation in the Inter-Group Prisoner's Dilemma*

Timothy N. Cason[a], Sau-Him Paul Lau[b], and Vai-Lam Mui[c]

[a]Department of Economics, Krannert School of Management, Purdue University, 403 W. State St., West Lafayette, IN 47907-2056, U.S.A. (corresponding author: cason@purdue.edu; +1 765 494 1737 (phone); +1 765 494 9658 (fax))

[b]Faculty of Business and Economics, University of Hong Kong, Pokfulam Road,
Hong Kong

[c]Department of Economics, Monash Business School, Monash University, P.O. Box 11E, Clayton,
Victoria 3800, Australia

October, 2018

Abstract

This paper studies theoretically and experimentally how success in prior interaction affects cooperation in the inter-group prisoner's dilemma (IPD). We develop a model of the IPD that incorporates group-contingent social preferences and bounded rationality to derive conditions under which an increase in pro-social concerns for an out-group will increase cooperation. We then report an experiment that shows the cooperation rate increases from 8 percent in a baseline one-shot IPD to 42 percent when the IPD is preceded by a coordination game played by members of the two groups. A post-experiment survey and chat coding results using a natural language classification game both show that successful prior interaction increases individuals' concerns for their out-group.

JEL Classification: C72, C92, D02

Key words: Prior Interaction, Identity, Inter-Group Prisoner's Dilemma

Acknowledgement: We are grateful to the Research Grant Council of Hong Kong (GRF747212) and the University Research Committee of the University of Hong Kong for financial support. For valuable comments and suggestions we thank George Akerlof, Yan Chen, Kim-Sau Chung, Phil Grossman, Dan Houser, John List, Birendra Rai, Bob Slonim, Matthias Sutter, Ted Turocy, seminar audiences at the Academia Sinica, Chapman, Monash, National Central, National Taipei, National Tsing Hua and Purdue Universities, Max Planck Institute for Tax Law and Public Finance, the Chinese University of Hong Kong, National University of Singapore, University of Innsbruck, University of Munich, University of Queensland, University of Tasmania, and conference participants at the Australasian Public Choice Conference, the Economic Science Association, and the Society for the Advancement of Behavioural Economics. Zhipeng Cai, Justin Krieg, and Kelvin Yuen provided valuable research assistance. We alone are responsible for any errors.

1. Introduction

Costly practices adopted by organizations, as well as observations by researchers of organization and community governance, suggest a widely-held belief that prior interactions can significantly affect inter-group cooperation. This paper studies theoretically and experimentally the hypothesis that prior interactions may influence individuals' concerns for the welfare of members of their out-group and affect cooperation in a one-shot inter-group prisoner's dilemma (hereafter IPD). Many interactions between groups resemble an IPD. The economics department and the business school of a university may need to decide whether to cooperate on a joint infrastructure project, such as an economics laboratory. Members of the marketing department and the engineering department of a firm may need to work together to develop a new product. Many environmental management problems require the cooperation of different groups. Whenever the material incentives are such that 'Defect' is the dominant strategy for both groups, but 'Both Cooperate' Pareto dominates 'Both Defect,' the interaction between the two groups is an IPD. Unfortunately, achieving cooperation in an IPD can be challenging. For example, Griffin and Hauser (1996) discuss how failures of cooperation between the marketing and the engineering departments in product development are common and can lead to significant losses for firms.

Consistent with the belief that successful prior interactions can promote cooperation in subsequent challenging interactions, some organizations actively invest in costly activities that promote social interactions for members from different units and divisions. HP and Tandem Computers pioneered the Silicon Valley Friday afternoon beer bust (Jacobson, 1998; Rao and Scaruffi, 2011), and HP also instituted daily company-wide coffee breaks to promote organization-wide social interactions (Rao and Scaruffi, 2011). Some organizations have their members participate in Outward Bound outdoor adventures that require intense team work (Knez and Camerer, 2000). Discussing the lessons from a large number of cases regarding inter-group collaboration in environmental management, Wondolleck and Yaffee (2000) observe that prior success in dealing with problems with strong common interest (Wondolleck and Yaffee, 2000, p.141) and informal inter-group social interactions such as field trips to conservation sites (Wondolleck and Yaffee, 2000, p.160-161) can provide an important foundation in dealing with more challenging problems facing groups.

Given the prominence of the conjecture that "prior interactions matter" in affecting inter-group cooperation, this paper investigates theoretically a possible microfoundation and experimentally tests whether it has empirical support. Specifically, inspired by the literature in psychology and economics on how identity and in-group out-group differences affect cooperation (Tajfel and Turner, 1979; Akerlof and Kranton, 2000; Chen and Li, 2009), this paper proposes and experimentally tests the *prior interaction hypothesis* for the IPD. This hypothesis states that successful prior inter-group interactions that produce rewards for members from different groups, even those that have no impact on the material payoff of a subsequent IPD played by these groups, can still increase cooperation in the IPD. This is because such successful prior interactions increase individuals' concerns for the welfare of their out-group and make cooperating with the out-group a more desirable action psychologically.

Building on Chen and Li (2009), we develop a group-contingent social preferences model for a (symmetric) IPD played by two *n*-player groups, in which every player is inequity averse (Fehr and Schmidt, 1999), but every player is more envious of or less charitable towards members of their out-group. Each group's decision is determined by majority rule. Not surprisingly, if social preferences are sufficiently strong, this IPD with group-contingent social preferences and pivotal voting has three equilibria: Everyone Cooperates, Everyone Defects, and a mixed-strategy equilibrium. The two pure-strategy equilibria that feature either full Cooperation or Defection are rarely observed in existing studies of IPDs (see, for example, Insko et al., 1990; Schopler et al., 2001; Halevy et al., 2008; Gong et al. 2009 and the references cited there). The mixed-strategy equilibrium, however, generates the counter-intuitive and implausible prediction that cooperation will decrease if individuals become more charitable or less envious of their out-group.

We then consider decision errors and study the Quantal Response Equilibria (QRE) of the IPD. We do this for the following reasons. Playing the IPD with majority and pivotal voting requires individuals to make non-trivial strategic calculations. When individuals are playing the IPD once and for the first time, they may make mistakes or have uncertain social preferences. Previous work has shown that the QRE can account for decision errors and stochastic preferences and can be consistent with observed behavior that is incompatible with counter-intuitive predictions of Nash equilibrium in many experimental games (see, for example, Goeree and Holt, 2001; Cason and Mui, 2005; Levine and Palfrey, 2007; Battaglini et al., 2010 and the references cited there).[1] In addition, the use of QRE enables us to utilize previous work on the QRE and equilibrium selection (McKelvey and Palfrey, 1995; Turocy, 2005) to provide conditions under which an increase in pro-social concerns for the other group increases cooperation in this model of the IPD that has multiple equilibria.

We also report a laboratory experiment to study empirically whether successful prior interaction increases individuals' concerns for their out-group and promotes cooperation in the IPD. The experiment implements a one-shot minimum effort coordination game as the prior interaction. Subjects are randomly assigned to different three-person groups, and they play an initial game to build group identity. We find that in a Baseline treatment in which two three-person groups play a one-shot IPD, only 8.3% of subjects cooperate. In the Inter-group Coordination treatment, the six members from two groups play a one-shot, six-person minimum effort coordination game prior to playing the one-shot IPD. Subjects achieve the

[1] The probabilistic choice of the QRE can be interpreted as reflecting decision errors or stochastic preferences (or preference "shocks"). These different interpretations result in the same mathematical model. For brevity, in the text we will typically refer to decision errors.

efficient outcome in all six-person coordination games, and this successful prior interaction increases subjects' cooperation rate in the IPD to 41.7%. A post-experiment survey and chat coding results of communication by subjects that use a natural language classification game (Houser and Xiao, 2011) both show that compared to the Baseline treatment, subjects in the Inter-group Coordination treatment show a stronger concern for the welfare of their out-group.

2. Related Literature

Following the seminal contribution by Akerlof and Kranton (2000, 2010), recent studies in economics have shown that identity induced in the laboratory can affect behavior. Researchers have found that common group identity increases contributions in public goods games (Eckel and Grossman 2005), facilitates coordination in the battle of sexes game (Charness et al., 2007) and the minimum effort game (Chen and Chen, 2011), and increases relation-specific investment (Morita and Servátka, 2013). Hargreaves Heap and Zizzo (2009) find that playing against trustors from the out-group reduces the return rates of trustees in a trust game. Chen and Li (2009) study how identity affects social preferences, and find that subjects are more envious of and less charitable to out-group members. Delaney and Jacobson (2014) consider a public good game model in which individual contributions to a public good, such as a dam, by in-group members benefit the in-group but hurt agents outside the group (the Outsiders). They report experimental evidence that the presence of negative downstream externalities reduces contributions by in-group members by half when they have closer contact with the Outsiders, but have no effects on in-group members' contributions when they have had no contact with the Outsiders.

The importance of group boundaries is emphasized in the emerging economics contributions on identity discussed above. Somewhat surprisingly, both the theoretical and experimental work discussed above focuses on how identity and group boundaries affect the strategic interactions in which all decision-makers are *individuals* (who may belong to different groups). Our paper, instead, considers how identity and group boundaries affect the strategic interactions in which each decision-maker is a *group*.

Our paper is also related to contributions that emphasize how social preferences can transform a prisoner's dilemma (PD) into a stag hunt game that has multiple equilibria in which Both Cooperate Pareto dominates Both Defect (see, for example, the early work of Sen, 1967; and Farrell and Rabin, 1996; Knez and Camerer, 2000; Ahn et al., 2001; Basu, 2010). More precisely, as Ellingsen et al. (2012) have pointed out, in the presence of social preferences it is possible that while the *game form* (which summarizes the objective features of strategies and payoffs) faced by the players is a prisoner's dilemma, the *game* (which involves von Neumann-Morgenstern utilities) being played is actually a stag hunt game.

In his study regarding how endogenous evolution of moral values affects cooperation, Tabellini (2008) also considers a one-shot individual PD in which agents care both about material payoff and the psychological utility from taking the morally correct action of cooperating. Once again, the psychological utility transforms the PD into a stag hunt game. Values evolve endogenously in Tabellini's model because parents make decisions that shape their child's concerns for moral satisfaction. Outside the context of the PD, Dufwenberg et al. (2016) show how social preferences can transform a land conflict social dilemma into a coordination game with Pareto-ranked equilibria. They present experimental evidence supporting this argument, and use this insight to construct a new policy to reduce land conflict.

All the work discussed above on the PD as a stag hunt game in utilities focuses on the individual PD. In contrast, we are interested in the inter-group PD. We report novel evidence that prior interaction increases individual's concerns for their out-group and promotes inter-group cooperation in the IPD. This evidence strengthens the case for studying how endogenous changes in social preferences can affect cooperation in PD-like situations.

Our work also relates to Sobel's (2005) observation about social preferences and repeated interactions. Sobel (2005, p. 420) argues that besides the familiar folk theorem and reputational arguments that emphasize, respectively, foregone future benefits in deterring cheating and players' uncertainty about their opponent's motive, there is a need to study a third mechanism regarding how repetition affects cooperation: "A history of positive interaction with someone leads you to care about that person's welfare." Sobel makes this point in the context of repeated play of the same stage game, while our study focuses on how prior success in a one-shot coordination game affects cooperation and concerns for the out-group in the one-shot IPD. Using a modified dictator game with an Individual-Team-Individual treatment and an Individual-Individual-Individual treatment, Crawford and Harris (2018) showed that interactions with other subjects when making decisions in the team dictator game affect subjects' preferences, as revealed by their choices in the individual dictator game that took place after the team dictator game.

This study also contributes to an emerging experimental economics literature that studies "sequential spill-over effects" in games, which investigates how the play of a first game may affect behavior in an unrelated second game. Knez and Camerer (2000) show that achieving the efficient outcome in a repeated seven-action minimum effort coordination game increases cooperation in a later three-action (multiple step) repeated PD compared to a control treatment. Ahn et al. (2001) report a similar result when the repeated coordination game is a two-action stag hunt game and the subsequent repeated game is the standard PD.[2] Similar sequential positive spill-over effects have also been found

[2] Ahn et al. (2001) also report a similar but quantitatively smaller result when each player first plays a series of stag hunt games with a different player each period under random matching, and then plays a series of PD with a different player each period under random matching.

involving other games (see, for example, Devetag, 2005; Brandts and Cooper, 2006; Cason et al., 2012).[3]

All these studies consider games in which every decision maker is an individual, and subjects play both the first game (the "prior interaction" in our terminology) and the second game (the "target interaction" in our terminology) multiple times, either in a fixed partner matching in which a subject interacts with the same subject, or in a random matching environment in which a subject interacts with a randomly chosen subject each time. None of these studies focus on whether the prior interaction increases individuals' concerns for members of their out-group. Our finding is novel as it shows that success in achieving the efficient outcome in a one-shot inter-group coordination game can increase cooperation in a subsequent one-shot IPD, through increasing individuals' concerns for the out-group.

Finally, this study adds to a surprisingly small economics literature on the IPD. A sizable literature in psychology has shown that groups cooperate less than individuals in the IPD (see, for example, Insko et al., 1990; Schopler et al., 2001; Gong et al. 2009 and the references cited there). This finding is crucial in leading to what psychologists refer to as the *discontinuity effect*, which states that "in mixed motive situations, inter-group interactions are more competitive or less cooperative than interindividual interactions" (Schopler et al., 2001, p. 632).

While many studies in economics investigate behavior in the individual PD, very few study the IPD. Charness and Sutter (2012) and Kugler et al. (2012) recently survey the fast-growing experimental literature on individual versus group decision making in economics. Virtually none of this work (covered either in Charness and Sutter (2012) or their on-line Appendix on Suggested Further Reading, or in Kugler et al. (2012)) studies the IPD.[4] Most studies in the small IPD literature focus on the repeated IPD (Bornstein et al., 1994; Insko et al., 1998; Goren and Bornstein, 2000; Kroll et al., 2013, Kagel and McGee, 2016; Cason and Mui, 2018). Halevy et al. (2008) extends the IPD to allow players to choose between contributing to helping in-group members and hurting out-group members. Our study is the first that links the recent literature of identity economics and the importance of prior interaction to the under-

[3] Researchers have also studied spill-over effects when subjects play two games simultaneously, see, for example, Bednar et al. (2012); Cason et al., (2012); and Falk et al. (2013), and the references cited there. Liu et al. (forthcoming) report an experiment in which each subjects plays a common historical game with two different matches for 100 rounds. After 100 rounds, the subject switches to a new game with one match but continues playing the historical game with the other match. They find behavioral spillover in their experiment.

[4] Charness et al. (2007) is one of the studies discussed by Charness and Sutter (2012) that investigates the PD. In that study, a subject plays the individual PD with another subject, but in one treatment a subject also gets one-third of the sum of the payoffs received by members of his/her (randomly induced) in-group. Almost all the work regarding the IPD discussed by Kugler et al. (2012) are from the psychology literature on the discontinuity effect. Exceptions include Morgan and Tindale (2002) (who consider PDs in the individual vs. individual condition, the group vs. group condition, and the individual vs. group condition) and Charness et al. (2007). Exploiting the fact that the Swiss military randomly assigns candidates for training program to different platoons, Goette et al. (2012) compares the behavior of individuals in such randomly assigned social groups to those of individuals in randomly assigned minimum groups. They find that the former cooperate more in the PD, but they consider the PD played by individuals. Chakravarty et al. (2016) study how Hindu and Muslim subjects in rural India play the individual stag hunt game and the individual PD differently with in-group and out-group members. In their experiment, subjects play the PD followed by the stag hunt game and the spill-over effect from one game to the other is not their focus.

studied IPD.

3. The Model

The purpose of this section is to study the comparative static question of how an increase in individuals' concerns for their out-group members' welfare affects the equilibrium cooperation rates in the IPD. We use a model of the IPD with social preferences and majority voting, and equilibrium selection arguments, to address these issues. The comparative static results will be used to interpret the experimental results.

3.1 The Inter-group Prisoner's Dilemma with group-contingent social preferences

Consider an IPD played by two groups. Each group consists of an odd number of $n = 2m + 1$ players, with m is a positive integer.[5] A group's decision is determined by majority voting, and members of each group cast their votes between Cooperate (C) and Defect (D) simultaneously. In making her decision, an individual needs to take into account how members from both her in-group and out-group will vote. Every member of a group always gets the same material payoff. In Table 1, the material payoff of each member of the two groups is given as a function of the decisions made by each group, through the majority voting rule. This game can be analyzed as a $2n$ players voting game.

We make the standard assumptions:

$$T > R > P > S \qquad (1)$$

$$2R > T + S \qquad (2)$$

Equation (1) guarantees that if players are only concerned about their material payoff, then a pivotal player will always vote for D and $\left(C, C\right)$ Pareto dominates $\left(D, D\right)$. Equation (2) implies that $\left(C, C\right)$ is the total-surplus maximizing outcome.

		Group 2	
		Cooperate	Defect
Group 1	Cooperate	*R, R*	*S, T*
	Defect	*T, S*	*P, P*

Table 1: The Inter-group Prisoner's Dilemma Game Form

[5] When n is even, we need to consider the implications of different tie-breaking rules (such as the flip of a coin, or a cooperate (or defect) default rule when tie occurs). For brevity, we do not consider even-sized groups here. Our results also apply for the model with $m = 0$. In this degenerate case, the IPD model becomes the individual PD. However, we do not examine the individual PD in this paper.

If individuals are only concerned about their material interests, then achieving cooperation in the one-shot IPD can be difficult. As discussed above, however, social preferences can transform a prisoner's dilemma into a stag hunt game that has multiple equilibria, including Both Cooperate and Both Defect.

There are many different forms of social preferences that can transform the IPD into a stag hunt game. The objective of our study is not to investigate which specific form of social preferences best explains behavior in the IPD. Our goal, instead, is to use a particular form of social preferences to illustrate the prior interaction hypothesis. For our purpose, we adopt the model of group-contingent social preferences that Chen and Li (2009) develop in their pioneering experimental work on identity and social preferences, as that model provides a natural way to capture the idea that individuals may have a stronger concern for the welfare of their in-group members than out-group members. In their model, individuals have identical, inequity-averse preferences (Fehr and Schmidt, 1999), and they report experimental evidence that subjects are more envious and less charitable to members of their out-group. In the following analysis, we assume that the $2n$ agents in the IPD have identical group-contingent preferences and are inequity averse. Individuals' preferences are given by the utility function:

$$v_j\left(\pi_j,\pi_k,n\right)=\pi_j-nw^O\left(n\right)\left(\pi_j-\pi_k\right);\ j,k=1,2;\ j\neq k\ , \tag{3}$$

where $v_j\left(.,.,.\right)$ is the utility of a player in group j, π_j is the material payoff received by a player in group j, π_k is the material payoff received by each player in group k, and $w^O\left(n\right)$ is the weight that a player in group j puts on the material payoff of each of the n players in her out-group. The reasons that v_j depends on π_j and π_k according to (3) are as follows.

Since members of a group in the IPD always receive the same material payoff, the effect of social preferences only arise from the possible difference in a player's payoff and those of her out-group members. We represent these features through the term $w^O\left(n\right)$ in (3), as

$$w^O\left(n\right)=\frac{1}{2n-1}\left(\rho r+\sigma s\right), \tag{4}$$

where

$$\rho>0>\sigma\ , \tag{5}$$

and $r=1$ if $\pi_j>\pi_k$, and $r=0$ otherwise. Similarly, $s=1$ if $\pi_j<\pi_k$, and $s=0$ otherwise. If an agent has a higher material payoff than an agent in her out-group, then the extent that she is *charitable* to an agent in her out-group is given by the charity parameter ρ. If an agent has a lower material payoff than an agent in her out-group, then the extent that she is *envious* of an agent in her out-group is given by the envy parameter σ. A rise in ρ and/or σ (which is negative) will lead to an increase in an agent's concerns for the out-group members.

Applying the social preferences specification (3) through (5) to the IPD in Table 1 results in the IPD with group-contingent social preferences (hereafter often simply referred to as the IPD) given in Table 2.

		Group 2	
		Cooperate	Defect
Group 1	Cooperate	R, R	$S+\frac{n}{2n-1}\sigma(T-S), T-\frac{n}{2n-1}\rho(T-S)$
	Defect	$T-\frac{n}{2n-1}\rho(T-S), S+\frac{n}{2n-1}\sigma(T-S)$	P, P

Table 2: The IPD with Group-Contingent Social Preferences

3.2 Nash equilibrium

We first focus on the (symmetric) Nash equilibria of the IPD. It turns out that the properties of the Nash equilibria, as well as the logit equilibria of the IPD (to be introduced in Section 3.3), depend on which one of the following four mutually-exclusive cases is satisfied:

$$R-\left[T-\frac{n}{2n-1}\rho(T-S)\right]\leq 0, \tag{6}$$

$$0<R-\left[T-\frac{n}{2n-1}\rho(T-S)\right]<P-\left[S+\frac{n}{2n-1}\sigma(T-S)\right], \tag{7}$$

$$0<P-\left[S+\frac{n}{2n-1}\sigma(T-S)\right]=R-\left[T-\frac{n}{2n-1}\rho(T-S)\right], \tag{8}$$

or

$$0<P-\left[S+\frac{n}{2n-1}\sigma(T-S)\right]<R-\left[T-\frac{n}{2n-1}\rho(T-S)\right]. \tag{9}$$

To interpret these conditions, consider the IPD with given material payoff parameters (P,R,S,T). When the value of ρ is sufficiently low and (6) is satisfied, the extent of social preferences is not strong enough that the prediction is similar to the pure self-interest model, which is a special case (with $\sigma=\rho=0$) of this IPD. In this case, a pivotal player will always choose D, and (D,D) is the unique pure-strategy equilibrium.[6]

When agents' concerns for the out-group increases—that is, when ρ and/or σ increase—we move from (6) to (7), (8), or (9). When (7), (8) or (9) is satisfied, if an agent is the pivotal decision maker of her group and the other group cooperates, then she strictly prefers to cooperate. While D will give her a higher material payoff, it will also lead her to suffer from a psychological disutility because her material payoff is higher than that of the members of the other group. When (7), (8) or (9) holds, this disutility is significant enough (with $R-\left[T-\frac{n}{2n-1}\rho(T-S)\right]>0$) to make D unattractive, and the IPD becomes a coordination game with Pareto ranked equilibria. Focusing on symmetric equilibria, this game has two pure-strategy Nash equilibria: Everyone Cooperates and Everyone Defects, and a symmetric mixed-strategy Nash equilibrium. The two pure-strategy equilibria predict either complete Cooperation or complete Defection, which are rarely observed in existing studies of IPDs (Insko et al., 1990, Schopler et al., 2001; Halevy et al., 2008; Gong et al. 2009).

On the other hand, cooperation in IPDs may be consistent with the mixed-strategy equilibrium. To derive the mixed-strategy Nash equilibrium, let $x\in\{0,1,\ldots,n\}$ denote the number of players who vote for C in a player's out-group. A player who is pivotal in her group will be willing to randomize iff the utility difference between choosing C and D is zero. Thus, the equilibrium probability of choosing C, q^*, is given by

$$\begin{aligned}&\left[\sum_{x=0}^{m}\binom{n}{x}\left(q^*\right)^x\left(1-q^*\right)^{n-x}\right]\left\{P-\left[S+\frac{n}{2n-1}\sigma(T-S)\right]\right\}\\&=\left[\sum_{x=m+1}^{n}\binom{n}{x}\left(q^*\right)^x\left(1-q^*\right)^{n-x}\right]\left\{R-\left[T-\frac{n}{2n-1}\rho(T-S)\right]\right\}\end{aligned}, \qquad (10)$$

[6] More precisely, we are actually focusing on trembling hand perfect Nash equilibrium in our analysis. As common in voting games, there also exists a pure-strategy Nash equilibrium Everyone Cooperates that involves the use of weakly-dominated strategy. In this equilibrium, because every player is non-pivotal, every player is indifferent between choosing the weakly dominated strategy C and D. Hence, Everyone Cooperates can be supported as a Nash equilibrium. This equilibrium, however, in not robust, as any positive probability of any member in his group choosing D will make a player strictly prefer D to C.

where $\sum_{x=0}^{m}\binom{n}{x}\left(q^{*}\right)^{x}\left(1-q^{*}\right)^{n-x}$ is the probability that an agent's out-group will defect (with *m* or less members of the group voting C), while $\sum_{x=m+1}^{n}\binom{n}{x}\left(q^{*}\right)^{x}\left(1-q^{*}\right)^{n-x}$ is the probability that her out-group will cooperate. It can be shown that the symmetric mixed-strategy equilibrium of the IPD is unique when condition (7), (8) or (9) holds. Furthermore, Proposition 1 states the counter-intuitive predictions of the mixed-strategy equilibrium of the IPD. (All proofs of the results in the main text are given in Appendix A).

Proposition 1.

At the unique symmetric mixed-strategy equilibrium of the IPD,

(a) $\frac{\partial q^{*}}{\partial \rho}<0$;

(b) $\frac{\partial q^{*}}{\partial \sigma}<0$.

Some aspects of Proposition 1 are illustrated in Figure 1, which plots the relationship of q^{*} versus the charitable parameter ρ in an IPD with $n = 3$. For this illustration, we fix the material payoff parameters as those in our experiment: $(P,R,S,T)=(54,132,28,162)$, and fix the value of σ at -0.112, as in Chen and Li (2009, Table 2). The two dotted lines separate the graph into three regions, corresponding to conditions (6), (7) and (9). It is observed that q^{*} is a decreasing function of ρ when (7), (8) or (9) is satisfied. In the IPD, an increase in agents' charitable concerns for others makes choosing C more attractive. If all players of the out-group continue to choose C with probability q^{*}, then the player will strictly prefer C. To ensure that the player will still randomize (being indifferent between C and D) after an increase in ρ, players of the out-group must now play C with a *lower* probability at the new mixed-strategy equilibrium. This result illustrates a well-known observation that because each player's probability of randomization is chosen to make other players willing to randomize in a mixed-strategy equilibrium, the mixed-strategy equilibrium often generates counter-intuitive predictions.

Summing up, while the mixed-strategy equilibrium in a model incorporating group-contingent social preferences alone can explain why cooperation can occur in the IPD, it generates the counter-intuitive and implausible prediction that greater concerns for the out-group members will decrease cooperation.

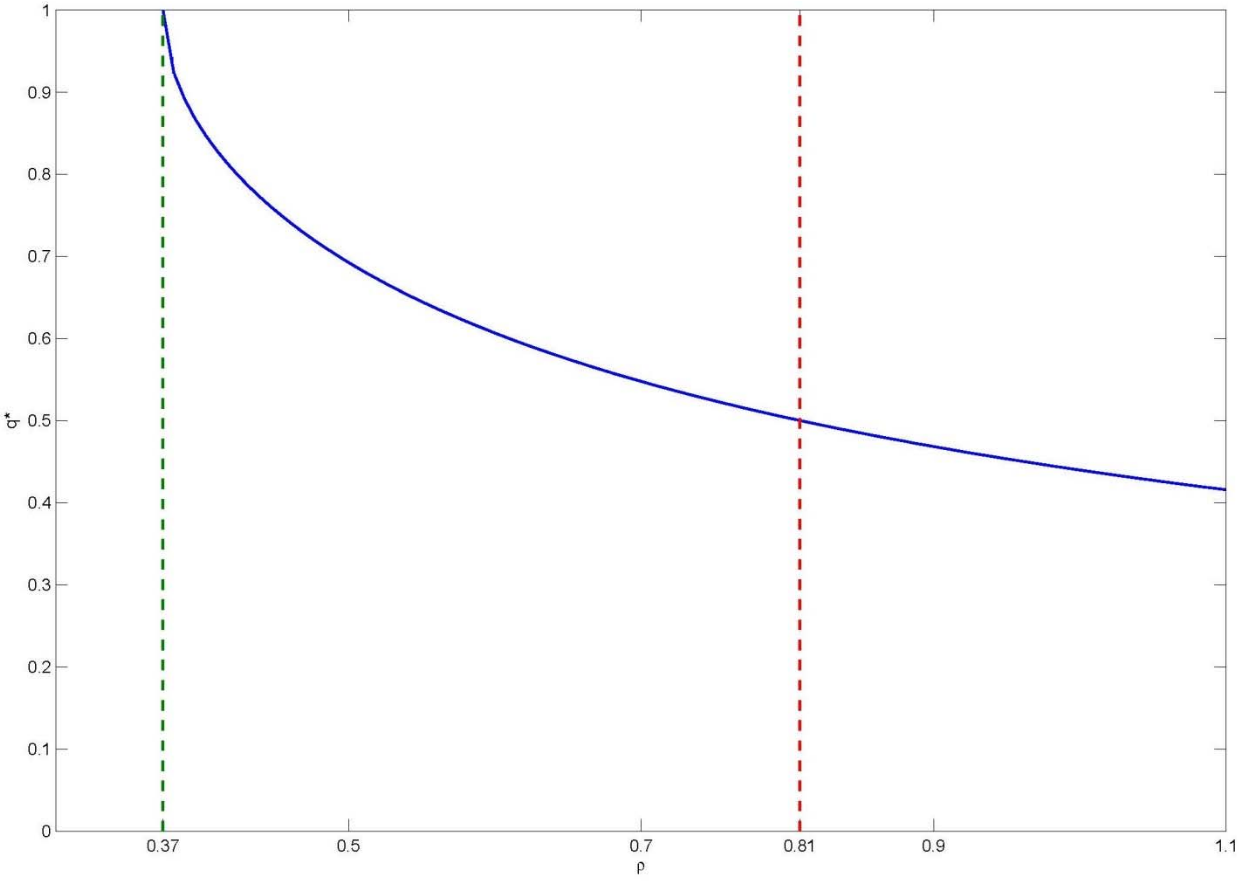


Figure 1: Symmetric Mixed strategy equilibrium

3.3 Quantal Response Equilibrium

Motivated by the observation that decision makers may make mistakes or experience preference shocks, especially in an unfamiliar strategic environment, McKelvey and Palfrey (1995) developed the concept of QRE. Subsequent work shows that the QRE can account for observed behavior that is incompatible with counter-intuitive predictions of Nash equilibrium in many experimental games (Goeree and Holt, 2001; Cason and Mui, 2005; Levine and Palfrey, 2007; Battaglini et al., 2010). Both decision errors and preference shocks can be important in the one-shot IPD in our experiment. We therefore consider the QRE of the IPD with group-contingent social preferences, and focus on the logistic quantal response function and the corresponding logit equilibrium.[7]

[7] McKelvey and Palfrey (1995) first consider the case that the players use a general quantal response function capturing decision errors, and define the QRE as the equilibrium when all players' quantal responses are mutually consistent. They then obtain further results when the players use a particular quantal response function, the logistic quantal response function (McKelvey and Palfrey, 1995, Section 3). The corresponding equilibrium when players use logistic quantal response function is called the logit equilibrium. The logit equilibrium has been widely used in the applications of the QRE, and is a special case of the regular QRE that ensures that the QRE has empirically falsifiable implications (see, Goeree et al., 2008, Haile et al., 2008, and Goeree et al.,

Let α_{ij} be the probability that agent i in group j will cooperate. Let $\left(\alpha_{1j},...,\alpha_{ij},...\alpha_{nj},\alpha_{1k},...,\alpha_{nk}\right)=\left(\alpha_{ij},\alpha_{-ij}\right)$ be the strategy profile adopted by the $2n$ agents, where α_{-ij} denote the strategy chosen by agents other than agent ij . The logistic quantal response function of player ij , $g_{ij}\left(\alpha_{-ij}\right)$, specifies player ij 's probability of playing C as a function of α_{-ij} and is defined as:

$$g_{ij}\left(\alpha_{-ij}\right)=\frac{e^{\lambda u_{ij}\left(1,\alpha_{-ij}\right)}}{e^{\lambda u_{ij}\left(1,\alpha_{-ij}\right)}+e^{\lambda u_{ij}\left(0,\alpha_{-ij}\right)}}=\frac{1}{1+e^{\lambda\left[u_{ij}\left(0,\alpha_{-ij}\right)-u_{ij}\left(1,\alpha_{-ij}\right)\right]}},\forall i=1,...,n,\forall j=1,2 \tag{11}$$

where $u_{ij}\left(1,\alpha_{-ij}\right)$ (resp. $u_{ij}\left(0,\alpha_{-ij}\right)$) is agent ij 's expected utility when she cooperates (resp. defects) and others play α_{-ij} . The logit precision parameter λ captures how sensitive an agent's decision is to the utility difference between playing C and D: $\lambda = 0$ implies that actions consist of all errors and the quantal response involves randomization with probability 0.5 between C and D, while $\lambda = \infty$ means that there is no error and she will choose the best response to others' strategies. Other than these extreme cases, the RHS of (11) implies that a player will *better respond*: she will choose both C and D with a positive probability, and the action that gives her a higher expected utility will be played with a higher probability. The logit equilibrium is a strategy profile $\left(\alpha_{ij}^{*},\alpha_{-ij}^{*}\right)$ satisfying the fixed point conditions:

$$\alpha_{ij}^{*}=g_{ij}\left(\alpha_{-ij}^{*}\right),\forall i=1,...,n,\forall j=1,2 \tag{12}$$

Focusing on the symmetric logit equilibrium such that every agent cooperates with the same probability ($\alpha_{ij}=\alpha,\forall i=1,...,n,\forall j=1,2$), the equilibrium is determined by:

$$\alpha^{*}=g(\alpha^{*},\lambda;\rho,\sigma;P,R,S,T), \tag{13}$$

where $g(\alpha,\lambda,\rho,\sigma,P,R,S,T)$ is the symmetric logistic quantal response function. For simplicity, we shall sometimes suppress the fact that the logistic quantal response function also depends on the social preferences parameters and the material payoff parameters,[8] and simply write the logistic quantal response function of the IPD, $g(\alpha,\lambda)$, as follows:

2016). For example, Chen and Chen (2011) adapt Chen and Li's (2009) group-contingent social preferences model to the minimum effort coordination game and study the logit equilibrium of the modified coordination game.

[8] Note that the symmetric logistic quantal response function $g(\alpha,\lambda,\rho,\sigma,P,R,S,T)$ is represented in various simpler forms in this paper, such as $g(\alpha,\lambda)$ in (14), as well as $g(\alpha,\lambda,\rho)$ and $g(\alpha,\lambda,\sigma)$ in (A14) in Appendix A, depending on the context.

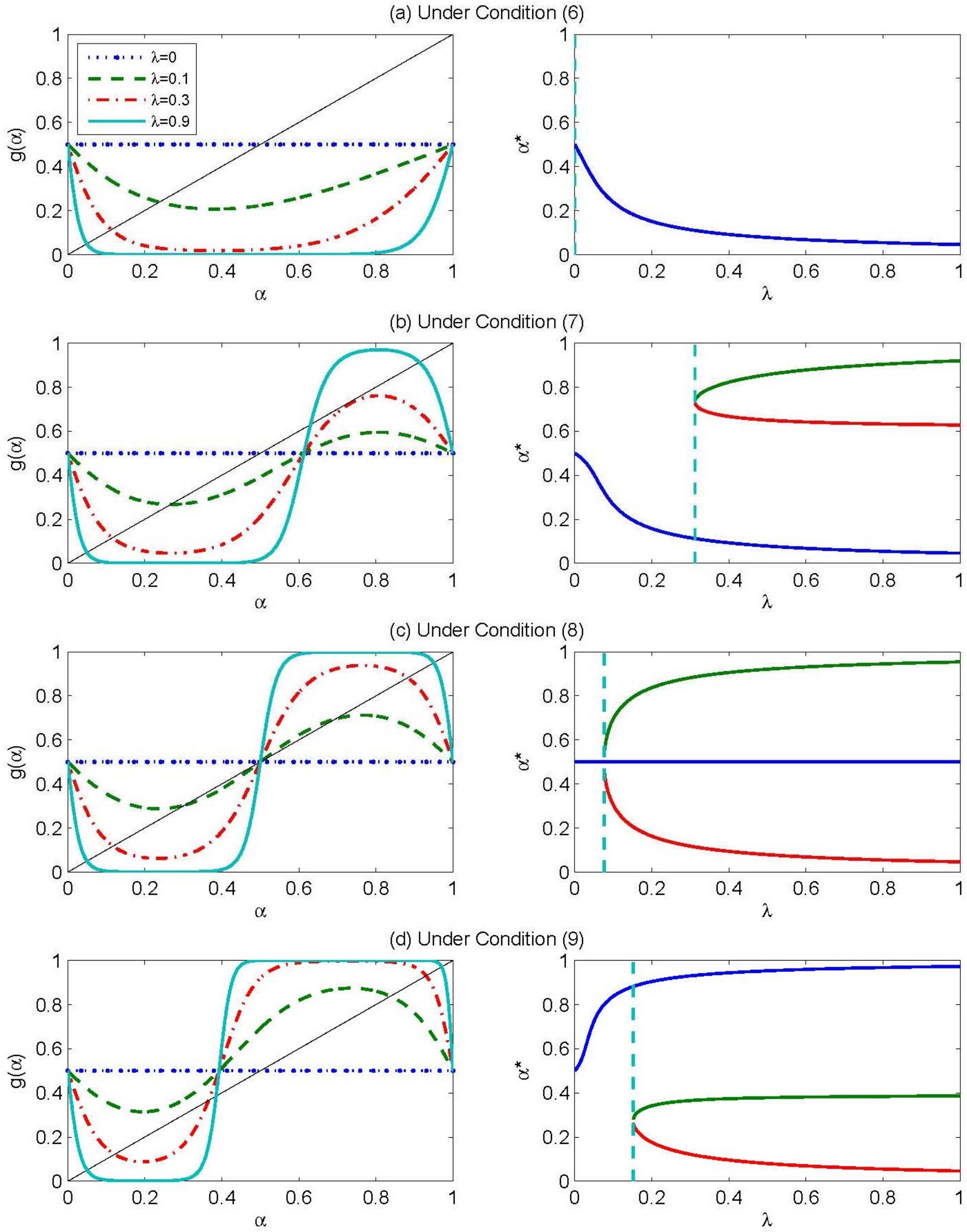
$$g(\alpha,\lambda)=\frac{1}{1+e^{\lambda\left[u_{ij}(0,\alpha,\ldots,\alpha)-u_{ij}(1,\alpha,\ldots,\alpha)\right]}}$$

$$=\frac{1}{1+e^{\lambda\left[\binom{2m}{m}\alpha^m(1-\alpha)^m\right]\left\{\left[\sum_{x=0}^{m}\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\right]\left[P-\left[S+\frac{n}{2n-1}\sigma(T-S)\right]\right]+\left[\sum_{x=m+1}^{n}\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\right]\left[\left[T-\frac{n}{2n-1}\rho(T-S)\right]-R\right]\right\}}} \qquad (14)$$

where $(1,\alpha,\ldots,\alpha)$ and $(0,\alpha,\ldots,\alpha)$ means player ij cooperates and defects respectively, when all other players cooperate with probability α.

To interpret (14), note that C and D will generate a different expected utility if and only if agent ij is the pivotal decision-maker in her group, that is, if and only m members choose C while the other m members choose D in her group. The probability that player ij is pivotal is $\binom{2m}{m}\alpha^m(1-\alpha)^m$. If the other group defects (which occurs with probability $\left[\sum_{x=0}^{m}\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\right]$), the utility difference of player ij between D and C is $P-\left[S+\frac{n}{2n-1}\sigma(T-S)\right]$. If the other group cooperates (which occurs with probability $\left[\sum_{x=m+1}^{n}\binom{n}{x}\alpha^x(1-\alpha)^{n-x}\right]$), the corresponding utility difference is $\left[T-\frac{n}{2n-1}\rho(T-S)\right]-R$. Therefore, the expression that appears after parameter λ in the RHS of (14) is simply the difference in agent ij's expected utilities generated by her actions D and C given the behavior of others.

Some key properties of the logistic quantal response function $g(\alpha,\lambda)$, which are important for subsequent results, are illustrated in Figure 2, using the same parameters as in Figure 1. First, when $\lambda = 0$, a player is completely insensitive to the differences in expected utility between playing C and D and will play each strategy with equal probability. Thus, $g(\alpha,0)=0.5$ for all $\alpha\in[0,1]$. Second, when $\lambda > 0$, the probability that a player is pivotal, $\binom{2m}{m}\alpha^m(1-\alpha)^m$, equals zero when $\alpha = 0$ or $\alpha = 1$. If everyone else always chooses C ($\alpha = 1$) or always chooses D ($\alpha = 0$), then a player will not be pivotal and hence will be indifferent between C and D. As a result, her quantal response will be given by $g(0,\lambda)=g(1,\lambda)=0.5$. Third, if condition (6) holds, then $g(\alpha,\lambda)<0.5$ for all $\alpha\in(0,1)$ and $\lambda > 0$. Otherwise, when $\lambda > 0$, $g(\alpha,\lambda)<0.5$ for $0<\alpha<q^*$, $g(q^*,\lambda)=0.5$, and $g(\alpha,\lambda)>0.5$ for $q^*<\alpha<1$, where q^* is defined in (10). The intuition is that if others cooperate with a probability less


(a) Under Condition (6)

(b) Under Condition (7)

(c) Under Condition (8)

(d) Under Condition (9)

Figure 2: Logit Equilibrium

(resp. larger) than q^*, then by definition of the mixed-strategy equilibrium, a player gets a higher (resp. lower) utility by playing D instead of C. Thus, a player's quantal response is to cooperate with a probability less (resp. larger) than 0.5. Fourth, an increase in λ causes $g(\alpha,\lambda)$ to shift downward for $0<\alpha<1$ when (6) holds, but causes $g(\alpha,\lambda)$ to shift downward for $\alpha<q^*$ and to shift upward for $\alpha>q^*$ when (7), (8) or (9) holds. Intuitively, as a player becomes more sensitive to the utility differences of the strategies, she will choose the better response with a higher probability.

Now we consider the logit equilibrium, which is determined by the intersection of the logistic quantal response function $g(\alpha,\lambda)$ and the 45-degree line. The properties regarding the logit equilibrium correspondence, $\alpha^*(\lambda)$, of the IPD are given in Proposition 2.[9]

Proposition 2.

(a) When condition (6) holds, the range of the logit equilibrium correspondence $\alpha^*(\lambda)$ is $[0,0.5]$.

(b) When condition (7) holds, the range of $\alpha^*(\lambda)$ is $[0,0.5]\cup[q^*,1]$.

(c) When condition (9) holds, the range of $\alpha^*(\lambda)$ is $[0,q^*]\cup[0.5,1]$.

Since the behavior of $g(\alpha,\lambda)$ differs for the 4 different conditions (6) to (9), it is not surprising that the logit equilibrium correspondence $\alpha^*(\lambda)=g(\alpha^*(\lambda),\lambda)$ also differs with respect to these regions. Proposition 2 states that $\alpha^*(\lambda)$ cannot exist in the interval $(0.5,1]$ under condition (6), in the interval $(0.5,q^*)$ under condition (7), and in the interval $(q^*,0.5)$ under condition (9). To see, for example, why $\alpha^*(\lambda)$ cannot exist in the interval $(0.5,q^*)$ under condition (7), note that under condition (7), for every $\alpha\in(0.5,q^*)$, the agent prefers playing D to playing C. Since an agent better responds under the logistic quantal response function, $g(\alpha,\lambda)<0.5<\alpha$, and $\alpha\in(0.5,q^*)$ cannot be a logit equilibrium. These results show that the logit equilibrium imposes refutable restrictions regarding the agents' behavior in the IPD.[10]

[9] Note that either condition (7) or (9) covers condition (8) in the limit. Thus, we do not state explicitly the range of $\alpha^*(\lambda)$ under condition (8) in Proposition 2.

[10] The theoretical propositions—like propositions regarding how parameters concerning risk attitudes can affect economic behavior—involve social preference parameters and the logit precision parameter that are not directly observable to researchers.

3.4 Principal path of the logit equilibrium correspondence

Proposition 2 provides testable implications regarding the equilibrium probability of playing C (α^*) in the IPD. However, the range of possible value of α^* is still quite large, since the precision parameter λ can take any non-negative value. The predictions of Proposition 2 are not very sharp, unless there are good reasons to pin down the value of λ. Another problem also arises even if the precision parameter can be narrowed down. As observed in McKelvey and Palfrey (1995) as well as the proof of Proposition 2, the logit equilibrium correspondence for sufficiently large values of λ are multi-valued. In order to derive sharper predictions, we need to develop arguments to select an equilibrium path.

McKelvey and Palfrey (1995) showed that the logit equilibrium correspondence is a singleton when λ is sufficiently small, but generally contains multiple values when λ becomes higher. They further show that the graph of the logit equilibrium correspondence contains a unique branch which starts at the centroid (with $\lambda = 0$, at which players' behavior is completely random) and converges to a unique Nash equilibrium. Turocy (2005) calls this unique branch the principal branch. The principal branch, α^*_{prin}, is defined by

$$g\left(\alpha^*_{prin}\left(\lambda\right),\lambda\right)=\alpha^*_{prin}\left(\lambda\right) \tag{15}$$

for all $\lambda \geq 0$, where the dependence of α^*_{prin} on λ is expressed explicitly.

In the following analysis, we focus on this unique principal branch for the following reasons. First, for any parameter profile $(\rho,\sigma;P,R,S,T)$ that describes the material payoffs and social preferences of the agents, when (7), (8), or (9) holds, the graph of the logit equilibrium correspondence has multiple branches.[11] However, all branches other than the principal branch only exist when λ is larger than a strictly positive threshold value. On the other hand, the principal branch is the only branch that will generate a prediction for any logit precision parameter $\lambda \geq 0$. Second, Turocy (2005, Theorem 7) showed that in a 2×2 game with two strict Nash equilibria, the principal branch of the logit equilibrium correspondence converges to the risk-dominant equilibrium when $\lambda \to \infty$. We shall show that a similar result holds in this IPD with group-contingent social preferences played by $2n$ players.[12] Earlier research regarding equilibrium selection for pure coordination games suggests that while no selection criterion can fully explain observed behavior, risk dominance does have significant explanatory power (Camerer, 2003,

Our study is not designed to estimate these parameters. If one is interested in estimating these parameters, one can consider a design in which subjects play a large number of IPDs with different material payoffs.

[11] McKelvey and Palfrey (1995, Theorem 2) showed that each branch of the graph of the logit equilibrium correspondence converges to a Nash equilibrium when $\lambda \to \infty$.

[12] In the IPD studied in this paper, if D (resp. C) is a player's best reply when every other player chooses C with probability 0.5, then Everyone Defects (resp. Everyone Cooperates) is the risk dominant equilibrium (Harsanyi and Selten, 1988).

chapter 7). Because of these two reasons, we focus on the principal path when the logit equilibrium correspondence has multiple branches.

Based on Proposition 2, we derive further results about the principal branch of the logit equilibrium correspondence of the IPD, including that it converges monotonically to the risk dominant outcome. This is given in the following proposition.

Proposition 3.

As λ increases from 0, the principal branch of the logit equilibrium correspondence of the IPD, which starts from $\alpha^*_{prin}(0)=0.5$,

(a) is always in the interval $[0,0.5]$, and decreases monotonically in λ and converges to the risk-dominant equilibrium that Everyone Defects (i.e., $\lim_{\lambda\to\infty}\alpha^*_{prin}(\lambda)=0$), when condition (6) or (7) holds;

(b) is always at $\alpha^*_{prin}=0.5=q^*$, when condition (8) holds;

(c) is always in the interval $[0.5,1]$, and increases monotonically in λ and converges to the risk-dominant equilibrium that Everyone Cooperates (i.e., $\lim_{\lambda\to\infty}\alpha^*_{prin}(\lambda)=1$), when condition (9) holds.

These results are illustrated in Figure 3(a). We again consider $(P,R,S,T)=(54,132,28,162)$, and σ fixed at the value of -0.112. The implication of Proposition 3 is that, if agents always play the IPD according to the principal branch of the logit equilibrium correspondence, then as the logit precision parameter λ increases, the probability of choosing C (α^*_{prin}) will become closer to the equilibrium rate of cooperation—which equals either zero or one—given by the risk-dominant equilibrium.

3.5 Comparative static results

We now derive comparative static predictions on the player's equilibrium probability of choosing C (α^*_{prin}) in the IPD with group-contingent social preferences and bounded rationality. For example, one may want to predict what happens to α^*_{prin}, when ρ increases (an increase in an agent's charitable concerns for out-group members). For these comparative static results we suppose that agents always play according to the principal branch of the logit equilibrium correspondence.

First, we consider the prediction of the model when the logit precision parameter (λ) remains unchanged. This is given by the following proposition.

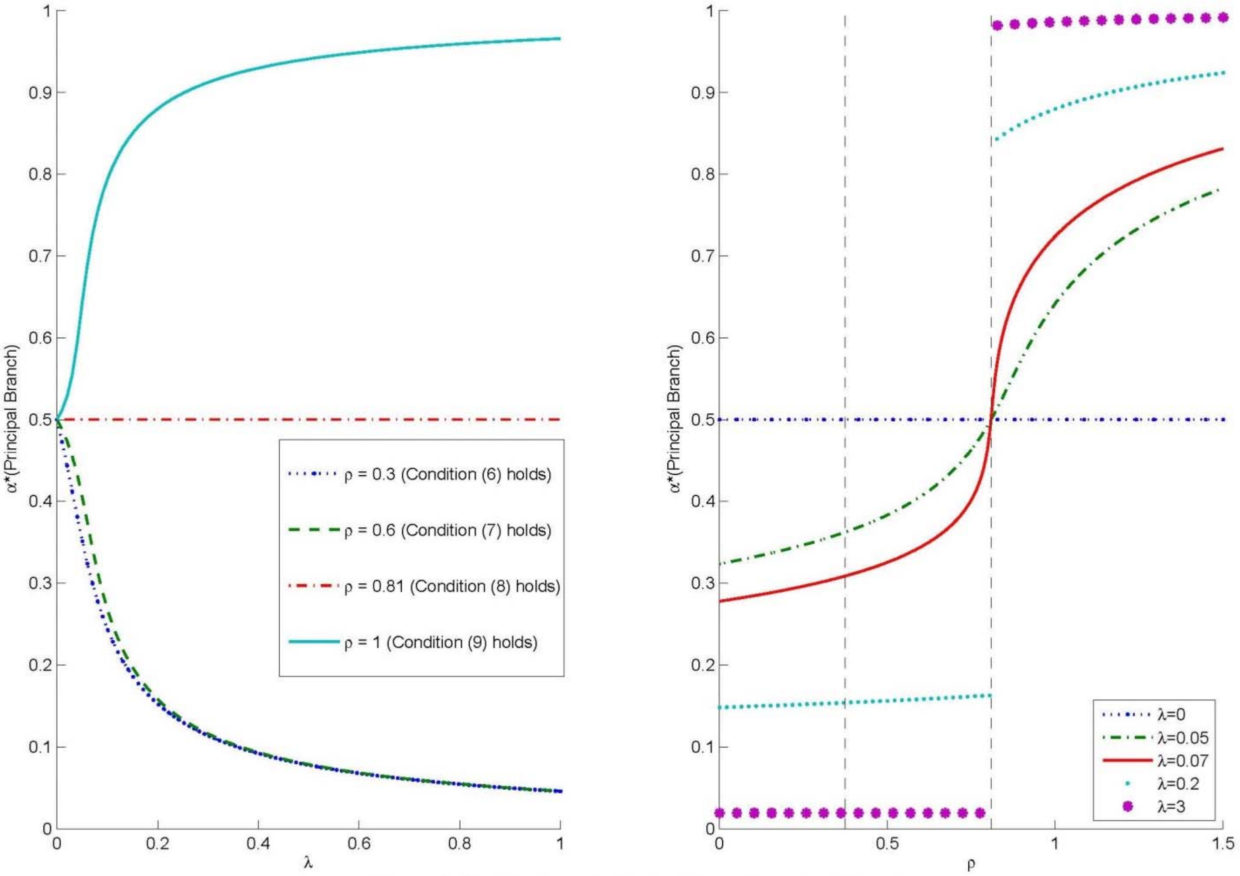


Figure 3: Equilibrium α (Principal Branch) against λ and ρ

Proposition 4.

An increase in the pro-social concerns for out-group members (i.e., either an increase in charitable or envious parameter) increases the players' probability of choosing C, at each level of the precision parameter. That is, for $\lambda > 0$,

$$\frac{\partial \alpha^*_{prin}}{\partial \rho} > 0\,; \tag{16}$$

$$\frac{\partial \alpha^*_{prin}}{\partial \sigma} > 0\,. \tag{17}$$

The monotonicity result in (16) can be observed in Figure 3(b), where we plot α^*_{prin} versus ρ , when all other parameters (including λ) remain unchanged for each path.

Proposition 4 implies that if a successful prior interaction makes individuals more charitable to or less envious of their out-group, other things being equal, it will increase individuals' probability of cooperation in equilibrium. Changing preferences towards the out-group has the following effects. First, an increase in an individual's concerns for her out-group increases the utility difference between Cooperate and Defect. As implied by the quantal response function in (14), even if other individuals' probability of cooperation does not change, at every level of the precision parameter (λ) the individual will now cooperate with a higher probability. Second, in equilibrium, an individual correctly expects that increased pro-social concerns cause both members of her group and her out-group to cooperate with a higher probability. That is, she "trusts" that both her members of her group and her out-group are more likely to cooperate.

In short, the successful prior interaction increases the equilibrium probability of cooperation because each individual is now responding to a higher probability of cooperation by all other players and is also picking her (better) response to this higher probability of cooperation on a "higher" quantal response function. The model shows that taking into account the implications of decision errors and preference shocks can be important in this environment when subjects only play the IPD once. In the presence of bounded rationality, if successful prior interaction increases individuals' concerns for the welfare of their out-group members, then individuals will expect (or "trust") that other players are more likely to cooperate in the IPD, and they will cooperate more in equilibrium. Our experimental design allows us to gather empirical evidence regarding how successful prior interaction affects the following three variables in the subsequent IPD: cooperation rates, individuals' concerns for the welfare of their out-group and their beliefs about how likely others will cooperate.

Proposition 4 is a comparative static result when λ is held constant. This result is relevant if a change in the parameter does not affect the logistic precision parameter. Now we turn to a stronger prediction of the model, irrespective of the level of λ . An exogenous change (such as through a treatment manipulation in an experiment) that is strong enough to change the IPD from the region of (6) or (7) to the region of (9), leads to the following corollary.

Corollary.

Suppose that there is an exogenous change in agents' pro-social concerns for the out-group that shifts the IPD under condition (6) or (7) to one under condition (9). This change causes the probability of choosing C to increase from a value below 0.5 to a value above 0.5, irrespective of the value of the precision parameter (λ) in both treatments, provided that $\lambda > 0$.

4. Experimental Design and Procedures

The experiment studies the IPD played by two groups of three members each, with the material payoffs given in Table 3.

		Group 2	
		Cooperate	Defect
Group 1	Cooperate	132, 132	28, 162
	Defect	162, 28	54, 54

Table 3: Material Payoffs (in HK$) of the IPD Experiment (HK$7.80≈US$1.00)

The experiment included three treatments. Twelve independent groups of six subjects participated in each treatment, for a total of 216 subjects. The timeline below summarizes each experimental treatment and highlights the differences between the treatments. Subjects read the instructions for a particular task (which were also read aloud by an experimenter) at the beginning of each task.

Experiments in psychology and economics have induced group identity in a variety of ways, such as through classification of artwork preferences (e.g., Chen and Li, 2009), or by allowing subjects to help in-group members in answering quiz questions (e.g., Morita and Servátka, 2013). In our study all treatments began with a simple minimum effort game (Van Huyck et al., 1990) to build initial group identity. As indicated in the experimental instructions in Appendix B, subjects could earn HK$19.50 by coordinating on the maximum integer (7) or as little as HK$10.50 by coordinating on the minimum integer (1). Since the goal of this task was to build group identity we wanted the subjects to be able to solve this coordination problem. Therefore, we allowed them to send a non-binding proposed choice followed by anonymous chat communication for two minutes before they were required to submit their final choice. This led to successful coordination on the Pareto optimal equilibrium of integer 7 by 65 of the 72 groups. Since essentially all subjects chose the highest number 7 on this coordination task, we do not have variation in coordination game behavior to relate to subsequent cooperation choices. Results and earnings from this preliminary game were displayed immediately to subjects.

In the Baseline treatment the groups then proceeded directly to the IPD, with payoffs (per player) shown in the right-hand panel of Table 3 paid in HK$. As in the minimum effort game, in the IPD considered in the Baseline and all other treatments, subjects first made a non-binding proposed choice and then engaged in a private, 3-player chat (this time for three minutes). The group's choice to defect or cooperate in the IPD was determined by majority vote. This design allows us to use the natural language classification game introduced in Houser and Xiao (2011) to examine the chats to investigate, among other issues, whether success in the prior interaction increases individuals' concerns for the welfare of their out-group members.

Before results of the IPD were shown, subjects submitted beliefs individually indicating their subjective likelihood that the other group voted in every possible way (3 Cooperate & 0 Defect, 2 C & 1 D, 1 C & 2 D, and 3 D). They were paid (up to HK$20) for accuracy based on a quadratic scoring rule.[13] Subjects also completed a simple risk assessment task and completed a post-experiment survey.

Baseline (in all treatments)	Added for Inter-group Coordination	Added for Inter-group Coordination+Communication
Task 1: 3-player minimum effort game (with chat) to build group identity		
	Task 2: 6-player minimum effort game (with chat among 6 players)	Task 2: 6-player minimum effort game (with chat among 6 players)
		3 chat communication phases before IPD: 3-player groups, all 6 players (both groups), again only 3-player groups
Task 3: Inter-group prisoner's dilemma played between 3-player groups (always preceded by 3-player chats)		
Task 4: Incentivized belief elicitation about other group's IPD voting		
Post-experiment survey, risk preference assessment and demographic questionnaire		

The other two treatments added an inter-group coordination task prior to the IPD, which may affect an agent's social preferences towards the other group. In the language of our model, the conjecture is that coordination success in this prior inter-group coordination game could raise ρ and/or σ. This coordination game was similar to the initial 3-player minimum effort coordination game, except this time played by all 6 players on both groups. We again allowed subjects to send a non-binding proposed choice and then permitted anonymous chat communication for two minutes before they submitted their final

[13] The quadratic scoring rule is incentive compatible (Savage, 1971), and since subjects did not learn about this task until after the IPD was completed, it could not have affected their choices in the IPD.

choice. This led to successful coordination on the Pareto optimal equilibrium of 7 by every one of the 36 groups. Results and earnings from this inter-group coordination game were displayed immediately to subjects. Importantly, this coordination game involving members from both groups occurs before subjects were informed that they would be playing an IPD. This kind of prior social interaction that took place *before* the game of interest differs from *in game* interactions that occurred after subjects have begun interacting in the game of interest—for example, communication between subjects when they are playing the prisoner's dilemma—that have been widely studied by economists.[14]

To focus on the "pure effect" of prior social interaction, our second treatment only added the coordination game prior to the IPD. If groups have the opportunity to interact prior to engaging in interactions resembling the IPD, however, they are also likely to have the opportunity to communicate with one another when they are playing the IPD. Our third treatment added communication between groups after they learned about the IPD. Following an initial 3-minute private chat among the 3 group members, a larger 6-player chat occurred for both groups, also for 3 minutes. This was followed by another private 3-player chat opportunity for group members and then the usual voting for cooperation or defection in the IPD.

Subjects were recruited from classes, through e-mail and posted announcements around the University of Hong Kong and they signed up using ORSEE (Greiner, 2015). No subject participated in more than one session. Two 6-person groups were conducted simultaneously, which helped maintain anonymity regarding group membership in the lab. The experimental software was written in zTree (Fischbacher, 2007). To ensure greater understanding about all aspects of the instructions and the tasks subjects had to complete, tasks 1, 2 and 3 included paid, computerized quizzes immediately following the reading of those instructions. Subjects were paid HK$2 for each correct quiz answer, and could earn up to HK$32 in total. (Any incorrect response provided a clarification for that question, based on text from the instructions.) Task understanding was excellent, since subjects scored 94.4% correct on average, and 106 of the 216 subjects scored 100% correct. Sessions required approximately 90 minutes to complete, and average earnings were HK$163.52 each, with an inter-quartile range of HK$115 to HK$231.

5. Experimental Results

Table 4 summarizes the experiment outcomes. Success in the inter-group coordination game has a large impact on individual decisions to cooperate in the IPD. This 5-fold increase in cooperation, from 8.3% to 44.4%, is highly statistically significant.[15] Without any inter-group interaction before the IPD, the

[14] Sally's (1995) survey discusses how changing different aspects of the social dilemma itself affects cooperation, with emphasis on the effects of adding communication to the social dilemma itself.

[15] Individual votes to defect or cooperate are not statistically independent, and votes across teams are also not independent in the inter-group coordination treatments because of the prior interaction in the coordination game. Therefore, to test for treatment

outcome (Defect, Defect) is by far the most common, but it decreases to only one-quarter of all outcomes in the inter-group coordination treatment. This change in group-level outcomes is also highly significant (Fisher's exact test *p*-value=0.012). Similarly, inter-group coordination significantly increases payoffs (*p*-value<0.01).

Adding inter-group communication raises cooperation above the level observed in the intermediate inter-group coordination treatment, both at the individual and group level (one-tailed *p*-values<0.05). As seen in the fifth row, the increase in the (Cooperate, Cooperate) group outcome frequency from 1 to 5 is also statistically significant, but only marginally for this sample size (Fisher's exact test one-tailed *p*-value=0.077). Profit earned from the IPD also increases significantly, from 88 to 107, when communication is introduced (one-tailed *p*-value<0.05).

Treatment:	Baseline	Intergroup Coordination	Intergroup Coordination+Communication
Individuals Voting to Cooperate	6/72 (8.3%)	32/72 (44.4%)	49/72 (68.1%)
Groups Cooperating	2/24 (8.3%)	10/24 (41.7%)	16/24 (66.7%)
(Defect, Defect) Outcomes	10	3	1
(Cooperate, Defect) Outcomes	2	8	6
(Cooperate, Cooperate) Outcomes	0	1	5
Average IPD Payoffs	60.8	87.8	107.0
Average Belief Other Group Cooperates	0.23	0.46	0.69

Table 4: Summary of Experiment Outcomes

Both the chat room content and the post-experiment survey provide supporting evidence that coordination success in the coordination game prior to the IPD increases the concerns for the welfare of the other group. In order to quantify the chat room content, we recruited an additional 34 University of Hong Kong undergraduate students from the same subject pool who had not participated in the earlier experiment. They attended one of two "coding sessions" with 17 subjects in each. These sessions implemented the natural language classification game introduced in Houser and Xiao (2011). In each session the subjects read the chat room communications from all three treatments and half of the

effects we estimate a probit model of the binary decision to cooperate, with estimated standard errors that are robust to unmodelled correlation across choices within sessions. Coefficient estimates on a dummy variable for the inter-group coordination treatment are highly significant (*p*-value<0.01). Other statistical tests reported in the text are based on similar modeling of the error structure to account for correlation, except for nonparametric tests which are only conducted on statistically-independent, session-level observations.

experimental sessions, and were asked to indicate whether certain goals or attitudes were expressed by group members. The coders also indicated what they thought the group would choose, and judged, based on chat communication, what the group believed their counterpart group would choose. We employed a coordination game in order to give subjects incentives to provide accurate evaluations of the qualitative chat data: In addition to a fixed participation payment these subjects earned up to HK$120 for six randomly drawn responses, through a HK$20 bonus for each question and chat room where their own classification matches the most popular classification in their session.

We assessed the reliability of this coding procedure using Cohen's Kappa (Krippendorff, 2003; Cohen, 1960). The most reliably coded information concerns predictions about whether groups will defect (which correlates almost perfectly with actual defection decisions), as well as the groups' beliefs about the cooperation choice of the other group (which is quite similar to the elicited beliefs summarized at the bottom of Table 4 above). This content analysis also reveals that groups usually predict correctly when their counterpart group will actually defect, with successful prediction rates of 66% in the coordination+communication treatment and 87% in the other two treatments.

Although it does not quite reach the "moderate" reliability threshold of 0.4 for Cohen's Kappa, we do see systematic variation across treatment in the response to the following coding question: "Did any member of this group indicate a goal of earning as much money as possible for all six players in the cluster group?" The coders indicated that this goal was expressed by 36% of the baseline groups, 45% of the inter-group coordination groups, and 57% of the coordination+communication groups. This provides support for the idea that successful prior interaction affects the players' objectives and attitudes towards the other group.

The post-experiment survey provides further evidence that the change in cooperation rates across treatments is due to changes in subjects' self-reported objectives. The top half of Table 5 shows that the fraction of subjects who stated an objective to earn as much as possible for their group decreased relative to the baseline when inter-group coordination or coordination+communication was introduced (*p*-value<0.01 for both pairwise comparisons). The fraction who stated an objective to earn as much as possible for all six people in the group increased across treatments (one-tailed *p*-value=0.028 for Baseline to Inter-group Coordination comparison; *p*-value<0.01 for Inter-group Coordination to Coordination+Communication comparison).

The comparative static results of our model suggest that if successful prior interaction increases individuals' concerns for their out-group, then individuals will expect that others are more likely to cooperate. Consistent with this prediction, beliefs also change across treatments in a systematic way, consistent with actual cooperation rates. The average belief that the other group will cooperate doubles from the baseline to the inter-group coordination treatment (*p*-value<0.01), and increases by the exact same percentage when adding communication (*p*-value<0.01).

Question: "In Task III, when you voted, how would you describe the strategies you used? Please select all that apply."

Treatment:	Baseline	Inter-group Coordination	Inter-group Coordination + Communication
"I tried to earn as much money as possible for me and my two teammates."	51/72 (70.8%)	38/72 (52.8%)	23/72 (31.9%)
"I tried to earn as much money as possible for all six people in my cluster group."	14/72 (19.4%)	25/72 (34.7%)	44/72 (61.1%)

Question: "Generally speaking, would you say the people can be trusted or that you can't be too careful in dealing with people?"

Fraction Responding "Usually not trusted"	Encountered a Cooperating Group	Encountered a Defecting Group
Member of a Cooperating Group	1/36 (2.8%)	11/48 (22.9%)
Member of a Defecting Group	17/48 (35.4%)	17/84 (20.2%)

Table 5: Selected Post-Experiment Questionnaire Responses

The post-experiment survey also included a standard "general trust" question from the World Values Survey (worldvaluessurvey.org), as shown in the bottom half of Table 5. A majority of subjects (150 out of 216) indicated that others can be "Usually trusted." The table shows an interesting pattern that emerged, however, among the 46 individuals who indicated that others can be "Usually not trusted." Not surprisingly, individuals who cooperated but encountered a defecting group were much more likely to indicate that others cannot be trusted (*p*-value<0.01). Members of defecting teams' responses were also correlated with the choice made by their paired group. Surprisingly, however, those who interacted with a cooperating group reported a *greater* lack of trust in others (*p*-value=0.038). We conjecture that this correlation may reflect that those who harm the other (cooperating) group are engaging in ex post rationalization of their defection choice.

6. Conclusions

Motivated by the widely-held belief that prior interactions can significantly affect inter-group cooperation, this paper develops a simple, tractable model of how changes in individuals' concerns for their out-group affect cooperation in the IPD. We then report novel experimental findings showing that success in a prior inter-group coordination game increases individuals' concerns for the welfare of their out-group, and increases cooperation and individuals' beliefs about how likely others will cooperate in a subsequent IPD.

Our focus in this paper is on the effects of success, rather than failure in prior interaction, on cooperation in the IPD. Understanding whether and how failed prior interaction may affect inter-group cooperation is also important. If failure in prior interaction indeed reduces future cooperation, then it raises the question of whether and when some *offsetting successful prior interactions* that occurred after the failed interaction but before the IPD can sufficiently increase individuals' concerns for their out-group to enable the interacting groups to achieve significant cooperation in the IPD. In a follow-up study (Cason et al., 2018) we consider IPD games played by groups of equal and unequal sizes, and find that in both cases, failure in prior interaction (in the form of a stochastic coordination game) reduces cooperation in a subsequent one-shot IPD.

Generalizing our model to inter-group interactions involving groups of different sizes allows us to investigate how group sizes interact with social preferences, pivotal voting in majority group decision-making, and decision errors to generate novel testable predictions. When social preferences exist, if a person votes to defect in an IPD, she should account for how many people she is potentially harming in her out-group. Furthermore, under majority rule, differences in group sizes also affect the probability that an individual will be pivotal in her in-group. We show that taking into account such considerations, conditional on experiencing successful prior interaction, the individuals in the smaller group will cooperate more than individuals in the larger group. On the other hand, conditional on experiencing failed prior interaction, individuals in the smaller group will cooperate less than individuals in the larger group. Our experimental results are consistent with these theoretical predictions. These findings, together with the findings reported in the current study, demonstrate how the tractable theoretical model developed here is a useful first step toward enriching the toolbox for studying how prior interaction affects inter-group cooperation in social dilemmas.

Following Chen and Li (2009), our current model assumes that all agents have identical social preferences. The assumption of homogenous preferences is a sensible starting point, but extending the model to allow for heterogeneous preferences can allow us to consider other important questions. For example, in their discussion of prior interactions and inter-group cooperation in environmental management, Wondolleck and Yaffee (2000) report many cases in which individuals involved in inter-group interactions have argued that successful prior interaction can promote "trust" toward out-group members.[16] Our theoretical model captures one way through which prior interaction can increase trust. If

[16] Reflecting on his interaction with another member on the Cameron County Agricultural Coexistence Committee formed by farmers, federal and states government officials and environmentalists, a US Fish and Wildlife Service refuge manager observes

successful prior interaction increases individuals' concerns for the welfare of their out-group, then our comparative static results show that in equilibrium, individuals will cooperate more in the IPD, and they also *believe* that members in their out-group will cooperate more.[17] Future work can, however, consider a richer and more realistic model in which some individuals have standard preferences that cannot be changed by any prior interaction, while other agents have social preferences that are affected by successful and unsuccessful prior interactions. It will be useful to use such a richer model to study the effects of prior interactions and how they can help individuals better assess the type of a specific individual in the out-group, which affects how individuals will "trust" whether particular members of their out-group will cooperate in later social dilemmas.

Studying the effects of prior interactions on inter-group cooperation requires consideration of the role of non-economic motivations and how their endogenous changes affect inter-group interactions. This study, as well as the related literature discussed in Section 2, illustrate that these issues can be analyzed using the standard tools of economics.[18] A deeper understanding of the implications of prior interactions will require iterative dialogues between theories, laboratory experiments, field experiments, and analysis of naturally occurring data. Such careful iterative dialogues should eventually generate useful insights for policy makers and organizational designers regarding the best possible forms of prior interactions that can increase inter-group cooperation for a specific target interaction given the existing context, and shed light on the resources needed for implementing the required prior interactions. Such research may also indicate whether the target interaction is too ambitious given the prior context and the resources available to cultivate prior interactions aiming at facilitating inter-group cooperation. If necessary, efforts can be devoted in formulating a more realistic immediate goal and finding ways to achieve it.

that "We had a few informal lunches together. We even went on a fishing trip together, me and a county agent. I saw that his personal goals and his professional goals were not that different than mine…You don't build trust until you actually get to know people a little bit (Wondolleck and Yaffee, 2000, p. 161).

[17] More precisely, Proposition 4 implies that changing preferences towards the out-group has the following effects. First, an increase in an individual's concerns for her out-group increases the utility difference between Cooperate and Defect. As implied by the quantal response function in (14), even if other individuals' probability of cooperation does not change, at every level of the precision parameter (λ) the individual will now cooperate with a higher probability. Second, in equilibrium, an individual correctly expects that increased pro-social concerns cause both members of her group and her out-group to cooperate with a higher probability. That is, she "trusts" that both her members of her group and her out-group are more likely to cooperate.

[18] See Sobel (2005) and Tabellini (2008) for two recent thoughtful discussions regarding how non-economic motivations and their endogenous changes can be fruitfully studied using the standard tool of economic theory.

References

Ahn, T.K., E. Ostrom, D. Schmidt, R. Shupp, and J. Walker, "Cooperation in PD Games: Fear, Greed, and History of Play," *Public Choice*, 106: 137–155, 2001.

Akerlof, G., and R. Kranton, "Economics and Identity," *Quarterly Journal of Economics*, 115: 715–753, 2000.

Akerlof, G., and R. Kranton, *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*, Princeton University Press: Princeton, 2010.

Ansell, C. and A. Gash, "Collaborative Governance in Theory and Practice," *Journal of Public Administration Research and Theory*, 18:543–571, 2008.

Basu, K., "The Moral Basis of Prosperity and Oppression: Altruism, Other-Regarding Behaviour and Identity," *Economics and Philosophy*, 26: 189-216, 2010.

Battaglini M., R. Morton, and T. Palfrey, "The Swing Voter's Curse in the Laboratory," *Review of Economic Studies*, 77: 61-89, 2010.

Bednar, J., Y. Chen, T.X. Liu, and S. Page, "Behavioral Spillovers and Cognitive Load in Multiple Games: An Experimental Study," *Games and Economic Behavior*, 74: 12-31, 2012.

Bornstein, G., I. Erev, and H. Goren, "The Effect of Repeated Play in the IPG and IPD Team Games," *Journal of Conflict Resolution*, 38: 690-707, 1994.

Brandts, J., and D. Cooper, "A Change Would Do You Good: An Experimental Study on How to Overcome Coordination Failure in Organizations, *American Economic Review*, 96: 669–693, 2006.

Camerer, C., *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton: Princeton University Press, 2003.

Cason, T., and V-L. Mui, "Uncertainty and Resistance to Reform in Laboratory Participation Games," *European Journal of Political Economy,* 21: 708-737, 2005.

Cason, T., and V-L. Mui, "Individual versus Group Choices of Repeated Game Strategies: A Strategy Method Approach," *Working Paper*, 2018.

Cason, T., Lau, P. and V-L. Mui, "The Impact of Group Size and Prior Interaction Failure on Cooperation in the Inter-group Prisoner's Dilemma," *Working Paper*, 2018.

Cason, T., A. Savikhin, and R. Sheremeta, "Behavioral Spillovers in Coordination Games," *European Economic Review*, 56: 233–245, 2012.

Chakravarty, S., M. A. Fonseca, S. Ghosh, and S. Marjit, "Religious Fragmentation, Social Identity and Cooperation: Evidence from an Artefactual Field Experiment in India," *European Economic Review*, 90: 265-279, 2016.

Charness, G. and M. Sutter, "Groups Make Better Self-Interested Decisions," *Journal of Economic Perspectives*, 26: 157-176, 2012.

Charness, G., L. Rigotti, and A. Rustichini, "Individual Behavior and Group Membership," *American Economic Review*, 97: 1340-1352, 2007.

Chen, R., and Y. Chen, "The Potential of Social Identity for Equilibrium Selection," *American Economic Review*, 101: 2562-2589, 2011.

Chen, Y., and S. X. Li, "Group Identity and Social Preferences," *American Economic Review*, 99: 431-457, 2009.

Cohen, J., "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20: 37-46, 1960.

Crawford, I., and D. Harris, "Social Interactions and the Influence of 'Extremists'," *Journal of Economic Behavior and Organization*, 153: 238-266, 2018.

Delaney J. and S. Jacobson, "Those Outsiders: How Downstream Externalities Affect Public Good Provision," *Journal of Environmental Economics and Management*, 67: 340–352, 2014.

Devetag, G., "Precedent Transfer in Coordination Games: An Experiment," *Economics Letters*, 89: 227–232, 2005.

Dufwenberg, M., Köhlin, G., Martinsson P., and H. Medhin, "Thanks but No Thanks: A New Policy to Reduce Land Conflict," *Journal of Environmental Economics and Management*, 77: 31–50, 2016.

Eckel, C., and P. Grossman, "Managing Diversity by Creating Team Identity," *Journal of Economic Behavior and Organization*, *58*: 371–392, 2005.

Ellingsen, T., M. Johannesson and J. Mollerstrom, "Social Framing Effects: Preferences or Beliefs," *Games and Economic Behavior*, 76: 117-130, 2012.

Falk, A., U. Fischbacher, and S. Gächter, "Living in Two Neighborhoods—Social Interaction Effects in the Laboratory," *Economic Inquiry*, 51: 563–578, 2013.

Farrell, J. and M. Rabin, "Cheap Talk," *Journal of Economic Perspectives*, 10: 103-118, 1996.

Fehr, E., and K. Schmidt, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114: 817-868, 1999.

Fischbacher, U., "z-Tree: Zurich Toolbox for Readymade Economic Experiments," *Experimental Economics*, 10: 171–8, 2007.

Goeree, J, and C. Holt, "Ten Little Treasures of Game Theory and Ten Intuitive Contradictions," *American Economic Review*, 91: 1402-1422, 2001.

Goeree, J, and C. Holt and T. Palfrey, "*Quantal Response Equilibrium,"* in S. Durlauf and L. Blume, eds., *New Palgrave Dictionary of Economics*, Palgrave Macmillan, 2008.

Goeree, J, and C. Holt and T. Palfrey, *Quantal Response Equilibrium: A Stochastic Theory of Games*, Princeton: Princeton University Press, 2016.

Goette, L., D. Huffman, and S. Meier, “The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups,” *American Economic Journal: Microeconomics,* 4: 101–115, 2012.

Gong, M., J. Baron, and H. Kunreuther, “Group Cooperation under Uncertainty,” *Journal of Risk and Uncertainty*, 39: 251-270, 2009.

Goren, H., and G. Bornstein, “The Effects of Intragroup Communication on Intergroup Cooperation in the Repeated Intergroup Prisoner’s Dilemma (IPD) Game,” *Journal of Conflict Resolution:* 44:700-719, 2000.

Greiner, B., “Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE,” *Journal of the Economic Science Association*, 1: 114-125, 2015.

Griffin, A., and J. Hauser, “Integrating R&D and Marketing: A Review and Analysis of the Literature,” *Journal of Product Innovation Management*, 13: 191–215, 1996.

Haile, P., A. Hortacsu, and G. Kosenok, “On the Empirical Content of Quantal Response Equilibrium,” *American Economic Review*, 98:180-200, 2008.

Halevy, N., Bornstein, G. and L. Sagiv, “‘Ingroup Love’ and ‘Outgroup Hate’ as Motives for Individual Participation in Intergroup Conflict: A New Game Paradigm,” *Psychological Science,* 19: 405-41, 2008.

Hargreaves Heap, S., and D. Zizzo, “The Value of Groups,” *American Economic Review*, 99: 295-323, 2009.

Harsanyi, J., and R. Selten, *A General Theory of Equilibrium Selection in Games*, Cambridge: MIT Press, 1988.

Houser, D., and E. Xiao, “Classification of Natural Language Messages Using a Coordination Game,” *Experimental Economics*, 14: 1-14, 2011.

Insko, C., J. Schopler, R. Hoyle, G. Dardis, and K. Graetz, “Individual-Group Discontinuity as a Function of Fear and Greed,” *Journal of Personality and Social Psychology*, 58: 68-79, 1990.

Insko, C., J. Schopler, M. Pemberton, J. Wieselquist, S. McIlraith, D. Currey and L. Gaertner, “Long-Term Outcome Maximization and the Reduction of Interindividual–Intergroup Discontinuity,” *Journal of Personality and Social Psychology,* 75: 695–710, 1998.

Jacobson, D., “Founding Fathers,” Stanford Magazine, July/August1998. Available at http://www.bandwidthco.com/history/computers/hp/Founding%20Fathers.pdf. Accessed 20 November, 2013.

Kagel, J., and P. McGee, ‘Team versus Individual Play in Finitely Repeated Prisoner Dilemma Games,’ *American Economic Journal: Microeconomics*, 8, 253-276, 2016.

Knez, M., and C. Camerer, “Increasing Cooperation in Prisoner’s Dilemmas by Establishing a Precedent of Efficiency in Coordination Games,” *Organizational Behavior and Human Decision Processes*, 82: 194-216, 2000.

Krippendorff, K., *Content Analysis: An Introduction to Its Methodology*, Sage Publications: Thousand Oaks, 2004.

Kroll, S., List, J., and C. Mason, "The Prisoner's Dilemma as Intergroup Game: An Experimental Investigation," in J. List and M. Price, eds., *Handbook on Experimental Economics and the Environment*, Edward Elgar: Cheltenham, 2013.

Kugler, T., Kausel, E., and Kocher, M., 'Are Groups More Rational than Individuals? A Review of Interactive Decision Making in Groups,' *Wiley Interdisciplinary Reviews: Cognitive Science*, 3: 471-482, 2012.

Levine, D., and T. Palfrey, "The Paradox of Voter Participation? A Laboratory Study," *American Political Science Review,* 101: 143-158, 2007.

Liu, T., Bednar, J., Chen, Y. and S. Page, "Directional Behavioral Spillover and Cognitive Load Effects in Multiple Repeated Games," forthcoming, *Experimental Economics*, doi.org/10.1007/s10683-018-9570-7.

McKelvey, R., and T. Palfrey, "Quantal Response Equilibrium for Normal Form Games," *Games and Economic Behavior*, 10: 6-38, 1995.

Morgan, P., and Tindale, R., 'Group versus Individual Performance in Mixed-Motive Situations: Exploring an Inconsistency,' *Organizational Behavior and Human Decision Processes*, 87: 44-65, 2002.

Morita, H., and M., Servátka, "Group Identity and Relation-Specific Investment: An Experimental Investigation," *European Economic Review*, 58: 95-109, 2013.

Rao, A., and P. Scaruffi, *A History of Silicon Valley: The Greatest Creation of Wealth in the History of the Planet,* Omniware: Palo Alto, 2011.

Sally, D., "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992," *Rationality and Society*, 7: 58-92, 1995.

Savage, L. J., "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66: 783-801, 1971.

Schopler, J., C. Insko, J. Wieselquist, M. Pemberton, B. Witcher, R. Kozar, C. Roddenberry and T. Wildschut, "When Groups are More Competitive than Individuals: The Domain of the Discontinuity Effect," *Journal of Personality and Social Psychology,* 80: 632–644, 2001.

Sen, A., "Isolation, Assurance, and the Social Rate of Discount," *Quarterly Journal of Economics*: 81: 112-124, 1967.

Sobel, J., "Interdependent Preferences and Reciprocity," *Journal of Economic Literature*, 43: 392–436, 2005.

Tabelllini, G., "The Scope of Cooperation: Values and Incentives," *Quarterly Journal of Economics*, 123: 905-950, 2008.

Tajfel, H., and J. Turner, "An Integrative Theory of Intergroup Conflict." In Stephen Worchel and William Austin, eds., *The Social Psychology of Intergroup Relations*, Monterey, CA: Brooks/Cole, 1979.

Turner, J., "Some Current Themes in Research on Social Identity and Self-Categorization Theories," in N. Ellemers, R. Spears, and B. Doosje, eds., *Social Identity: Context, Commitment, Content*, Oxford: Blackwell, 1999.

Turocy, T., "A Dynamic Homotopy Interpretation of the Logistic Quantal Response Equilibrium Correspondence," *Games and Economic Behavior*, 51: 243-263, 2005.

Van Huyck, J., R. Battalio, and R. Beil, "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure," *American Economic Review*, 80: 234-48, 1990.

Wondollock, J. and S. Yaffee, *Making Collaboration Work: Lessons from Innovation in Natural Resource Management*, Washington D.C.: Island Press, 2000.

Appendix A: Proofs (Not for Publication)

(A) Proof of Proposition 1

We first show that that there exists a unique $q^* \in (0,1)$ such that (10) holds. Define

$$p_D^O(\alpha) = \sum_{x=0}^{m} \binom{n}{x} \alpha^x (1-\alpha)^{n-x}, \tag{A1}$$

which is the probability that the out-group votes to defect. From (A1), we have

$$\begin{aligned}\frac{\partial p_D^O(\alpha)}{\partial \alpha} &= -n(1-\alpha)^{n-1} + \sum_{x=1}^{m} \binom{n}{x} \left[x\alpha^{x-1}(1-\alpha)^{n-x} - (n-x)\alpha^x (1-\alpha)^{n-x-1} \right] \\ &= -n(1-\alpha)^{n-1} + \left[\binom{n}{1}(1-\alpha)^{n-1} + \sum_{x=2}^{m} \binom{n}{x} x\alpha^{x-1}(1-\alpha)^{n-x} \right] - \sum_{x=1}^{m} \binom{n}{x}(n-x)\alpha^x (1-\alpha)^{n-x-1}.\end{aligned}$$

It is easy to see that

$$\sum_{x=2}^{m} \frac{n!}{(x-1)!(n-x)!} \alpha^{x-1}(1-\alpha)^{n-x} = \sum_{j=1}^{m-1} \frac{n!}{j!(n-j-1)!} \alpha^j (1-\alpha)^{n-j-1}.$$

Therefore, for $\alpha \in [0,1]$,

$$\frac{\partial p_D^O(\alpha)}{\partial \alpha} = \frac{-n!}{m!(n-m-1)!} \alpha^m (1-\alpha)^{n-m-1} = \frac{-n!}{m!m!} \alpha^m (1-\alpha)^m < 0. \tag{A2}$$

Re-write (10) as

$$p_D^O(q^*) = \frac{\left\{ R - \left[T - \frac{n}{2n-1}\rho(T-S) \right] \right\}}{\left\{ P - \left[S + \frac{n}{2n-1}\sigma(T-S) \right] \right\} + \left\{ R - \left[T - \frac{n}{2n-1}\rho(T-S) \right] \right\}}. \tag{A3}$$

When (7), (8) or (9) holds, $R - \left[T - \frac{n}{2n-1}\rho(T-S) \right] > 0$ and $P - \left[S + \frac{n}{2n-1}\sigma(T-S) \right] > 0$. Therefore, the RHS of (A3) is positive and less than 1. From (A1) and (A2), $p_D^O(0) = 1$, $p_D^O(1) = 0$, and $p_D^O(\alpha)$ is continuous and strictly decreasing for $\alpha \in (0,1)$. The Intermediate Value Theorem implies that there exists a unique $q^* \in (0,1)$ such that (A3), or equivalently (10), holds.

From (A3), it is straightforward to show that $p_D^O(q^*)$ is strictly increasing in ρ and in σ. Since $p_D^O(\alpha)$ is strictly decreasing in α according to (A2), we obtain (a) and (b).

(B) Proof of Proposition 2

Define the weighted utility difference between D and C if the out-group defects as

$$ud_D(\alpha) = p_D^O(\alpha)\left\{P - \left[S + \frac{n}{2n-1}\sigma(T-S)\right]\right\}, \tag{A4}$$

and the weighted utility difference between C and D if the out-group cooperates as

$$ud_C(\alpha) = \left[1 - p_D^O(\alpha)\right]\left\{R - \left[T - \frac{n}{2n-1}\rho(T-S)\right]\right\}. \tag{A5}$$

Note that (A2) implies that $\frac{\partial ud_D(\alpha)}{\partial \alpha} < 0$ and $\frac{\partial ud_C(\alpha)}{\partial \alpha} > 0$ for $\alpha \in (0,1)$.

Re-write $g(\alpha,\lambda)$ in (14) as

$$g(\alpha,\lambda) = \frac{1}{1 + e^{\lambda\binom{2m}{m}\alpha^m(1-\alpha)^m\left[ud_D(\alpha) - ud_C(\alpha)\right]}}, \tag{A6}$$

or

$$\ln\left[\frac{1 - g(\alpha,\lambda)}{g(\alpha,\lambda)}\right] = \lambda\binom{2m}{m}\alpha^m(1-\alpha)^m\left[ud_D(\alpha) - ud_C(\alpha)\right]. \tag{A7}$$

According to (13) and (A7), we know that $\alpha^* \in [0,1]$ is a logit equilibrium of the IPD iff for some $\lambda \geq 0$, α^* satisfies

$$\ln\left(\frac{1-\alpha^*}{\alpha^*}\right) = \lambda\binom{2m}{m}\left(\alpha^*\right)^m\left(1-\alpha^*\right)^m\left[ud_D(\alpha^*) - ud_C(\alpha^*)\right]. \tag{A8}$$

It is obvious that

$$\ln\left(\frac{1-\alpha^*}{\alpha^*}\right)\begin{cases} > 0 & when \quad 0 < \alpha^* < 0.5 \\ = 0 & when \quad \alpha^* = 0.5 \\ < 0 & when \quad 0.5 < \alpha^* < 1 \end{cases}. \tag{A9}$$

Now, consider the RHS of (A8). Under (6), $R - \left[T - \frac{n}{2n-1}\rho(T-S)\right] \leq 0$ and thus, $ud_D(\alpha^*) - ud_C(\alpha^*) > 0$. Combining with (A9), there exists a $\lambda \geq 0$ such that (A8) holds iff $\alpha^* \in (0,0.5]$. The range of α^* is $(0,0.5]$.

Under (7) or (9), $R-\left[T-\frac{n}{2n-1}\rho\left(T-S\right)\right]>0$ and q^* exists in the interval $\left(0,1\right)$. Since $\frac{\partial ud_D(\alpha)}{\partial\alpha}-\frac{\partial ud_C(\alpha)}{\partial\alpha}<0$, it can be shown that for $\lambda>0$,

$$ud_D(\alpha)-ud_C(\alpha)\begin{cases} > ud_D(q^*)-ud_C(q^*)=0 & when \quad 0<\alpha<q^* \\ = ud_D(q^*)-ud_C(q^*)=0 & when \quad \alpha=q^* \\ < ud_D(q^*)-ud_C(q^*)=0 & when \quad q^*<\alpha<1 \end{cases}. \tag{A10}$$

Furthermore, it can be shown that $q^*>0.5$ when (7) holds, as follows. When n is odd, the $n+1$ binomial coefficients are symmetric. Therefore,

$$p_D^O(0.5)=0.5^n\sum_{x=0}^{m}\binom{n}{x}=0.5^n\left[\frac{1}{2}\sum_{x=0}^{n}\binom{n}{x}\right]=0.5,$$

where the last equality follows from $\sum_{x=0}^{n}\binom{n}{x}=2^n$. When (7) holds, the RHS of (A3) is less than 0.5. It follows from (A2) that $q^*>0.5$. Combining the above results, we conclude that the range of α^* is $\left(0,0.5\right]\cup\left(q^*,1\right)$.

Similarly, when (9) holds, the RHS of (A3) is larger than 0.5 and $q^*<0.5$. Moreover, the range of α^* is $\left(0,q^*\right)\cup\left[0.5,1\right)$.

Finally, it can be shown that the end points, $\left\{0,q^*,1\right\}$, of the above half-open or open intervals correspond to $\lambda=\infty$.[19] This proves Proposition 2.

[19] We can interpret λ in (A8) as an inverse function of α^*. Under (6), (7) or (9), when $\alpha^*\to 0$ from above, $\lim_{\alpha^*\searrow 0}\lambda=\frac{1}{\binom{2m}{m}\left[ud_D(0)-ud_C(0)\right]}\frac{\lim_{\alpha^*\searrow 0}\left[\ln\left(\left(1-\alpha^*\right)/\alpha^*\right)\right]}{\lim_{\alpha^*\searrow 0}\left[\left(\alpha^*\right)^m\left(1-\alpha^*\right)^m\right]}=+\infty$. Under (7) or (9), when $\alpha^*\to 1$ from below, $\lim_{\alpha^*\nearrow 1}\lambda=\frac{1}{\binom{2m}{m}\left[ud_D(1)-ud_C(1)\right]}\frac{\lim_{\alpha^*\nearrow 1}\left[\ln\left(\left(1-\alpha^*\right)/\alpha^*\right)\right]}{\lim_{\alpha^*\nearrow 1}\left[\left(\alpha^*\right)^m\left(1-\alpha^*\right)^m\right]}=+\infty$. Under (7), the range of the logit equilibrium correspondence $\alpha^*\left(\lambda\right)$ contains $\left(q^*,1\right)$ but not $\left(0.5,q^*\right)$, so α^* can only tend to q^* from above. We have $\lim_{\alpha^*\searrow q^*}\lambda=\frac{\ln\left(\left(1-q^*\right)/q^*\right)}{\binom{2m}{m}\left[\left(q^*\right)^m\left(1-q^*\right)^m\right]}\frac{1}{\lim_{\alpha^*\searrow q^*}\left[ud_D(\alpha^*)-ud_C(\alpha^*)\right]}=+\infty$. Under (9), the range of the logit equilibrium correspondence $\alpha^*\left(\lambda\right)$ contains $\left(0,q^*\right)$ but not $\left(q^*,0.5\right)$, so α^* can only tend to q^* from below. We have $\lim_{\alpha^*\nearrow q^*}\lambda=\frac{\ln\left(\left(1-q^*\right)/q^*\right)}{\binom{2m}{m}\left[\left(q^*\right)^m\left(1-q^*\right)^m\right]}\frac{1}{\lim_{\alpha^*\nearrow q^*}\left[ud_D(\alpha^*)-ud_C(\alpha^*)\right]}=+\infty$. In all cases, the values of λ and α^* satisfy (A8).

(C) Proof of Proposition 3

Under (6), the range of $\alpha^*(\lambda)$ is $[0,0.5]$. It is obvious that $\alpha^*_{prin} \in [0,0.5]$.

Under (7), the range of $\alpha^*(\lambda)$ is $[0,0.5] \cup [q^*,1]$. Theorem 3 of McKelvey and Palfrey (1995) established that $\alpha^*(\lambda)$ is upper hemicontinuous. They also showed in their Lemma 1 that there exists $\tilde{\lambda} > 0$, such that for all $\lambda \in [0,\tilde{\lambda}]$, the logit equilibrium is unique. These properties, together with the initial point $\alpha^*_{prin}(0) = 0.5$, imply that $\alpha^*_{prin}(\lambda) \in [0,0.5]$ for every $\lambda \in [0,\tilde{\lambda}]$. Next, consider what happens starting from $\alpha^*_{prin}(\tilde{\lambda})$, which is in the interval $[0,0.5]$. We can use similar arguments as above to show that for every $\lambda \geq \tilde{\lambda}$, $\alpha^*_{prin}(\lambda) \in [0,0.5]$.[20]

To consider the monotonicity part, we first derive a useful result. Differentiating (A6) with respect to λ, we obtain

$$sign\left[\frac{\partial g(\alpha,\lambda)}{\partial \lambda}\right] = sign\left[ud_C(\alpha) - ud_D(\alpha)\right]. \tag{A11}$$

Combining (A10) with (A11), we conclude that

$$\frac{\partial g(q^*,\lambda)}{\partial \lambda} \begin{cases} <0 & for \quad \alpha \in (0,1) \quad under \quad (6) \\ <0 & for \quad 0 < \alpha < q^* \quad under \quad (7) \quad or \quad (9) \\ >0 & for \quad q^* < \alpha < 1 \quad under \quad (7) \quad or \quad (9) \end{cases}. \tag{A12}$$

Under (6) or (7), pick $\lambda_1 > 0$ and the corresponding point $(\lambda_1, \alpha^*_{prin}(\lambda_1))$ in the graph of α^*_{prin}. Then consider $\lambda_2 > \lambda_1$. Using (15) and (A12), we have $g(\alpha^*_{prin}(\lambda_1),\lambda_2) - \alpha^*_{prin}(\lambda_1) < g(\alpha^*_{prin}(\lambda_1),\lambda_1) - \alpha^*_{prin}(\lambda_1) = 0$. On the other hand, we have $g(0,\lambda_2) - 0 = 0.5 - 0 > 0$. It follows from the Intermediate Value Theorem that there exists an $\alpha^*_{prin}(\lambda_2) \in (0, \alpha^*_{prin}(\lambda_1))$ such that $g(\alpha^*_{prin}(\lambda_2),\lambda_2) - \alpha^*_{prin}(\lambda_2) = 0$. Thus, $\alpha^*_{prin}(\lambda)$ decreases monotonically in λ. Since the only Nash equilibrium in the interval $[0,0.5]$ is $\alpha^* = 0$ (i.e., Defect), we conclude that $\lim_{\lambda\to\infty} \alpha^*_{prin}(\lambda) = 0$. This proves (a).

When (8) holds, $q^* = 0.5$. We conclude from $g(q^*,\lambda) = 0.5$ and (15) that $\alpha^*_{prin}(\lambda) = 0.5$ always. This proves (b).

[20] In essence, the "gap" $(0.5, q^*)$ in the range of the (upper hemicontinuous) logit equilibrium correspondence under (7) leads to the result that the principal branch that starts from the centroid ($\alpha^*_{prin}(0) = 0.5$) cannot escape from $[0,0.5]$.

Under (9), the range of $\alpha^*(\lambda)$ is $\left[0,q^*\right]\cup[0.5,1]$, and there is a gap $\left(q^*,0.5\right)$ in the range of $\alpha^*(\lambda)$. Combining with the initial point $\alpha^*_{prin}(0)=0.5$, we can use the same arguments as above to show that $\alpha^*_{prin}(\lambda)\in[0.5,1]$. Consider $\lambda_2>\lambda_1>0$. Using (15) and (A12), we have $g\left(\alpha^*_{prin}(\lambda_1),\lambda_2\right)-\alpha^*_{prin}(\lambda_1)>g\left(\alpha^*_{prin}(\lambda_1),\lambda_1\right)-\alpha^*_{prin}(\lambda_1)=0$ under (9). We also have $g(1,\lambda_2)-1=0.5-1<0$. It follows from the Intermediate Value Theorem that there exists an $\alpha^*_{prin}(\lambda_2)\in\left(\alpha^*_{prin}(\lambda_1),1\right)$ such that $g\left(\alpha^*_{prin}(\lambda_2),\lambda_2\right)-\alpha^*_{prin}(\lambda_2)=0$. Therefore, $\alpha^*_{prin}(\lambda)$ increases monotonically in λ. Since the only Nash equilibrium in the interval $[0.5,1]$ is $\alpha^*=1$ (i.e., Cooperate), we conclude that $\lim_{\lambda\to\infty}\alpha^*_{prin}(\lambda)=1$. This proves (c).

(D) Proof of Proposition 4

We first show that when (6), (7) or (9) holds,

$$\frac{\partial g\left(\alpha^*_{prin}(\lambda),\lambda\right)}{\partial\alpha}<1. \qquad \text{(A13)}$$

When (6) or (7) holds, we know from Lemma 1 that $g(0,\lambda)-0=0.5>0$ and $g(0.5,\lambda)-0.5<0$. Since the unique value of $\alpha^*_{prin}(\lambda)$ is defined by the intersection of the logistic quantal response function $g(\alpha,\lambda)$ and the 45-degree line in the interval $[0,0.5]$, we conclude that $g(\alpha,\lambda)$ intersects the 45-degree line from above. Thus, (A13) holds.[21]

When (9) holds, we know from Lemma 1 that $g(0.5,\lambda)-0.5>0$ and $g(1,\lambda)-1=-0.5<0$. Since the unique value of $\alpha^*_{prin}(\lambda)$ is defined by the intersection of

[21] A more formal proof is as follows. When (6) or (7) holds, $\alpha^*_{prin}(\lambda)$ is defined by the unique fixed point of $g(\alpha,\lambda)$ in the interval $[0,0.5]$. That is, $g\left(\alpha^*_{prin},\lambda\right)=\alpha^*_{prin}$, where $0\le\alpha^*_{prin}\le0.5$. We prove by contradiction. Suppose that $\frac{\partial g\left(\alpha^*_{prin}(\lambda),\lambda\right)}{\partial\alpha}\ge1$, then there exists an $\alpha_1<\alpha^*_{prin}$ in the neighbourhood of α^*_{prin} such that

$$\begin{aligned}&\frac{g\left(\alpha^*_{prin}\right)-g(\alpha_1)}{\alpha^*_{prin}-\alpha_1}\ge1\\ &\to g\left(\alpha^*_{prin}\right)-g(\alpha_1)\ge\alpha^*_{prin}-\alpha_1\\ &\to\alpha^*_{prin}-g(\alpha_1)\ge\alpha^*_{prin}-\alpha_1\\ &\to\alpha_1\ge g(\alpha_1).\end{aligned}$$

Together with $g(0)=0.5>0$, we conclude that there exists another fixed point $\alpha^*_{another}\in(0,\alpha_1]$ such that $g\left(\alpha^*_{another},\lambda\right)=\alpha^*_{another}$. This contradicts with the fact that α^*_{prin} is uniquely defined in $[0,0.5]$.

$g(\alpha,\lambda)$ and the 45-degree line in the interval $[0.5,1]$, we conclude that $g(\alpha,\lambda)$ intersects the 45-degree line from above. Thus, (A13) holds.[22]

According to (A13), $1-\frac{\partial g\left(\alpha^*_{prin}(\lambda),\lambda\right)}{\partial\alpha}\neq 0$. Therefore, the Implicit Function Theorem applies to (A13), which we re-write as $g\left(\alpha^*_{prin}(\lambda,\rho,\sigma),\lambda,\rho,\sigma\right)=\alpha^*_{prin}(\lambda,\rho,\sigma)$ to make the dependence of $\alpha^*_{prin}(\lambda,\rho,\sigma)$ on (ρ,σ) explicit. Furthermore, we have

$$\frac{\partial\alpha^*_{prin}(\lambda,\theta)}{\partial\theta}=\frac{\dfrac{\partial g\left(\alpha^*_{prin}(\lambda,\theta),\lambda,\theta\right)}{\partial\theta}}{1-\dfrac{\partial g\left(\alpha^*_{prin}(\lambda,\theta),\lambda,\theta\right)}{\partial\alpha}};\theta=\rho,\sigma\,. \tag{A14}$$

Using (A4) to (A6), it is straightforward to show that $\frac{\partial g(\alpha,\lambda,\rho)}{\partial\rho}>0$ and $\frac{\partial g(\alpha,\lambda,\sigma)}{\partial\sigma}>0$. Combining these results with (A13) and (A14), we obtain (16) and (17).

[22] A formal proof similar to that in the previous footnote can be constructed.